DOCUMENT RESUME

ED 151 408 TM 006 919

AUTHOR Schutz, Robert W.
TITLE Specific Problems in the Measurement of Change: Longitudinal Studies, Difference Scores, and Multivariate Analyses. Quantification Laboratory Technical Report No. 27.
PUB DATE May 77
NOTE 33p.; Paper presented at the Annual Conference of the North American Society for the Psychology of Sport and Physical Activity (Ithaca, New York, May, 1977)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Achievement Gains; Analysis of Variance; Attrition (Research Studies); *Behavior Change; Cross Sectional Studies; Data Analysis; *Longitudinal Studies; *Measurement Techniques; Post Testing; Pretesting; Reliability; *Research Design; Sampling; Scores; *Statistical Analysis; Testing Problems; Validity

ABSTRACT

The measurement of change is such a broad topic that this article must limit its focus to a few specific subtopics. These specific topics include: longitudinal research design, attrition in research studies, the statistical analysis of difference scores, and the comparison of analysis of variance (ANOVA) and multivariate analysis of variance (MANOVA) techniques in analyzing repeated measures data. The purpose and sampling techniques, as well as the internal and external validity are discussed for each measurement technique. The author concludes that a considerable number of problems are inherent in the measurement and analyses of change, especially in research designs of a longitudinal nature. However, most of these problems can be avoided with sufficient care and planning prior to initiating the research project. The cross-sectional sequential type designs which are required for valid measures of developmental change are very costly--but necessary if the research is to have any scientific value. Multivariate statistical procedures utilizing complete data sets will provide for valid and relatively powerful tests of hypotheses. (Author/MV)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. W. Schutz

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED151408

— TM006 919 —

SPECIFIC PROBLEMS IN THE MEASUREMENT OF CHANGE:

LONGITUDINAL STUDIES, DIFFERENCE SCORES, AND MULTIVARIATE ANALYSES

Presented
at the

North American Society for the Psychology
of Sport and Physical Activity
Annual Conference

Ithaca, New York
May, 1977

Robert W. Schutz
Quantification Laboratory
School of Physical Education
and Recreation
University of British Columbia
Vancouver, B.C.

Quantification Laboratory
Technical Report #27
May, 1977

The general topic to which this paper is directed, the measurement of change, is so broad, so all-encompassing, that any one article or presentation must limit its focus on a few specific subtopics. A psychometrician or researcher interested in statistics and methodology dealing with change would need to use an extensive list of key words in his literature search to keep abreast in the field. Terms such as developmental, longitudinal, growth, trend, repeated measures, change, curve fitting, and stochastic are just some of the descriptors - each one of which constitutes a specialized and relatively extensive body of knowledge. The specific topics dealt with in this paper are longitudinal research design, the statistical analysis of difference scores, and the comparison between ANOVA and MANOVA techniques in analyzing repeated measures data.

LONGITUDINAL RESEARCH DESIGN

History

Cross-sectional studies, often with inadequate design and control, or longitudinal studies on a few subjects, were the most common basis for drawing inferences regarding developmental change in the early 1900's. Large scale longitudinal studies, like those of Terman, et al. (1925), were few and far between. However, by the 1950's and 60's, there seemed to be a considerable increase in the extent to which researchers were willing to embark on such extensive research projects, partially because of a better understanding of the pitfalls in cross-sectional research and also because of an increase in the availability of funding for this type of developmental research. More recently, methodological papers by Baltes (1968), Labouvie, Bartsch, Nesselroade, and Bates (1974), and Schaie (1965) have revealed numerous shortcomings in the basic research design commonly employed in longitudinal studies. In light of these papers, both the external and internal validity of many longitudinal studies must now be questioned.

Baltes (1968, p. 149), in discussing the traditional cross-sectional and longitudinal designs, states that, "In the light of present standards of research methodology, both research designs appear to be relatively naive," and then later (1965, p. 153) claims that, "... both conventional designs have such a total absence of control as to be of almost no scientific value." For those investigators

about to initiate such a research venture, there are procedures available which help circumvent these problems, but the researcher into the second or third year of a 20-year longitudinal study is now faced with a difficult decision. He must either abandon the study, and start again, or attempt to incorporate controls into the study in an attempt to establish as much external and internal validity as possible. Obviously, future longitudinal studies must be designed with considerable care, with an associated increase in labour, subject and financial costs. A brief description of the problems associated with longitudinal studies follows, along with possible solutions to these problems, as well as comparisons between longitudinal and cross-sectional designs.

Cross-Sectional Studies

Purpose and sampling. As Schaie (1973, p. 164) points out, valid research design and sound data collection methodology can be employed only if, "the specific developmental question is made explicit." If the sole purpose is to examine differences among cohorts at a single point in time, then a cross-sectional design will suffice. A major problem, however, is to obtain comparable samples from the different age groups. If one was to sample 20-year olds and 50-year olds from a given community, there would undoubtedly be a number of variables, other than age, distinguishing the two groups. The adventurous, or the dull, or the very talented, may have left the community, thus the 50-year olds are a particular residual group. Furthermore, if the sample was drawn from volunteers, not only would the external validity be limited, but so would the internal validity. It is unlikely that volunteers from a 20-year old population differ from 20-year old non-volunteers in the same manner and degree as 50-year old volunteers differ from their non-volunteer cohorts. Random sampling will permit comparisons among cohort populations within the sample domain, but inferences cannot be made beyond this population.

Internal and external validity. Cross-sectional studies confound the effects of aging with generational effects, thus introducing a source of error which may impair the internal validity of this design. The frequently held belief that many behavioral attributes decline with age, after peaking around age 25 was based on evidence gathered with cross-sectional studies. Subsequent longitudinal studies (Schaie and Strother, 1968; Schaie, Labouvie, and Barrett, 1973) have negated this hypothesis by showing virtually no change within

individuals up to ages 40 and 50, but considerable between cohort differences. Thus the early cross-sectional studies reflected between generation differences and yet were interpreted as differences due to aging.

The problem of non-random population attrition, called selective survival by Baltes (1968), affects the external validity of both cross-sectional and longitudinal designs. Evidence is cited (Baltes, 1968) that a specific population at say age 20, changes in its composition over time in a selective manner, so that the survivors by age 50 are the subjects who were the taller, and more intelligent ones in the original sample. With a cross-sectional design, there is no way to control and/or examine this phenomena.

Design and analysis. The usual experimental design for a cross sectional study is a single factor, randomized groups design. Appropriate analysis for a single dependent variable would be a one way ANOVA, with orthogonal polynomial decomposition of the sum of squares for cohorts possible if a trend analysis is desired. However, unlike a repeated measures design where distinct and appropriate error terms are available for each trend component, this design yields only a within-groups mean square which must be used as the denominator in all F tests. Consequently, the design results in statistical tests of relatively low power, both for the main effect and any single degree-of-freedom contrasts.

Longitudinal Studies

Purpose and sampling. The usual purpose of a longitudinal study is to examine changes within individuals in terms of physical or behavioral development. Consequently, the procedures of the past have involved obtaining a relatively large random sample at one point in time followed by repeated observations of the same subjects for a period of time (a few months up to a life time). As only one cohort is needed, the sampling procedures do not have the problems cross-sectional studies do in equating samples across cohorts. If the longitudinal study is going to be a life time study, or any considerable length of time, then a large initial base is necessary as considerable attrition is likely to occur (which causes numerous other problems). One major problem resulting from the continuous tracking and measuring of a large number of subjects is the financial cost - a NIH longitudinal study which monitored 50,000 children from

pregnancy to seven years old cost 60 million dollars (Wall and Williams, 1970). Obviously any large scale longitudinal study requires funding from wealthy foundations or governmental agencies.

Internal and external validity. In contrast to the cross-sectional studies which confound age and generation effects, the longitudinal study confounds the effects of aging with those related to cultural changes. Over a 20-year period many behavioral attributes show a pronounced change within society in general, and thus show up as a change within individuals. Society's attitudes towards working mothers or pornography, which have undoubtedly changed over the last 30 years, would show up as a change in attitude from age 20 to age 50 in a longitudinal study. Phenomena such as those cited here would probably be correctly interpreted; however, with many other variables it is questionable whether any change can be primarily accounted for by aging or by cultural changes.

Selective sampling, selective survival, and selective drop-out (terms from Baltes, 1968), all tend to lower the external validity of longitudinal studies. The population which is apt to volunteer for a longitudinal study tends to be of a higher socio-economic status and intelligence than non-volunteers (Rose, 1965), and attrition from such studies is also selective in that those subjects dropping out (both refusers and movers) tend to be of lower intelligence (Labouvie, et al., 1974; Schaie, et al., 1973).

A third problem associated with longitudinal studies, and one that is not present in cross-sectional studies, is the repeated testing effect. Labouvie, et al. (1974, p. 282) conclude that, "... the findings indicate that age-related longitudinal increases on intelligence variables are mainly due to retest effects." They feel that the internal validity of simple longitudinal studies is lowered to such an extent by repeated testing effects that any inferences about age-related changes are "unjustified and grossly misleading". There are two ways an investigator can test for and/or control for this testing effect. Schaie (1973) suggests retesting a subsample within a relatively short period of time, before any age or environmental influences are likely to have taken place, and if there is no change at this time, then the researcher can be confident that any differences in a year will not be due to the testing effect. If there are differences, then it will be necessary to utilize the other procedure, the introduction of a control group, which is discussed in the section on mixed designs.

Design and analysis. The most common method of analyzing longitudinal data is to treat it as a single factor, repeated measures design (or, if two or more groups, a k x p factorial experiment with repeated measures on the second factor, where k is the number of groups, and p is the number of testing sessions). The nature and degree of change over time can then be tested for statistical significance with either a MANOVA or a repeated measures ANOVA - the advantages and disadvantages of these two methods is discussed in a subsequent section. Trend analysis, a very powerful statistical test with a repeated measures design, provides an indication of the significance of any polynomial trends over testing sessions.

Bentler (1973), Nunnally (1967), and others advocate the use of a factor analytic technique to analyze longitudinal data. This procedure transposes the subject by test data matrix into a testing session by subject matrix and factor analyzes that, giving factors of people, each subject having a loading on each factor. If there were three factors, this would represent three different patterns of change over time. The problem with such analyses is that it rests on the assumption that individual differences in change can be grouped into types. It is this investigator's opinion that most differences in change over time among individuals are a matter of degree, not of type. Consequently, the factor solutions would not be very distinct.

Other less common procedures such as progressive partialing analysis (Nunnully, 1967), stochastic processes (Schutz, 1970), and time series (Gottman, McFall and Barnett, 1969) have potential as valuable statistical tools in explaining variability in patterns of change.

Mixed Longitudinal Cross-Sectional Designs

It has been shown that the two commonly used designs in studying developmental change both confound a component with the effects of age, longitudinal studies confound age and environmental or cultural effects, and cross-sectional studies confound age with generation differences. A third design is the time-lag study in which one age group is examined longitudinally, that is, a different sample of say 10 year olds are selected and tested every 5 years. This design then, while not even accounting for age, confounds generational and cultural effects. The obvious solution is to combine all three designs in an attempt to remove the

confounding effects. Schaie (1965) attempts to do this with his trifactor developmental model - a sequential research design which attempts to separate the effects of age, cohort, and time of measurement. The age effect indicates maturation of the individual, cohort effects should indicate hereditary effects, and time of measurement effects are indicative of changes due to environmental effects (although Baltes, 1968, suggests that the cohort component may also include environmental effects). Table 1 represents this sequential design. Note that the three rows represent three longitudinal studies, the columns represent cross-sectional designs (although only column 1960 samples all four cohort groups) and each of the four diagonals represent time-lag studies. Schaie formulates three equations, based on the premise that differences between cross-sectional measures, between longitudinal measures, and between time-lag measures, are each a sum of the two components which are confounded in these designs. Through a process of subtraction, he can then get independent estimates of each of the three components, age, cohort, and time. Such a procedure, however, requires six subsamples in order to get these three independent estimates. The design represented in Table 1 would therefore not be sufficient, and would require cohorts at 1970 and 1980, with testing continuing to the year 2010 in order to get complete 30-year longitudinal data on 6 cohort groups.

[Insert Table 1 about here]

Baltes (1968), while acknowledging Schaie's contribution to methodology in developmental research design, raises two objections to the trifactor model. The first objection, certainly a valid one, is that the three components, age, cohort and time, are not really mutually independent. Any one component can be replaced by a linear combination of the other two, thus giving rise to Baltes' (1968) bifactor model of age and cohort. The second objection raised by Baltes concerns Schaie's definition of the variation accounted for by the time of measurement component. The effects of maturation and environment cannot be isolated through direct measurement, causing the time component to be a confounded variable itself.

Using Baltes' bifactor model as the best available research design for development studies results in a classical p x q factorial design with repeated measures on the second factor (p being the number of cohort groups, and q the number of different age classifications under which each cohort group is tested).

Such a design can be analyzed by the repeated measures analysis of variance given in Table 2. This allows for an analysis of the age effect, the cohort

[Insert Table 2 about here]

effect, as well as the interaction which tests if the change over age is constant across the various cohort levels. Further polynomial breakdown on both the age and the cohort main effects are possible.

The bifactor and trifactor models of Baltes and Schaie, although accounting for age and cohort differences, still do not control for one of the major sources of invalidity in longitudinal studies, namely the effect of repeated testing. Both investigators, however, have made suggestions for testing and/or controlling for this effect. Essentially, these controls entail a separate control group for each cohort and age level. Thus, if the original cohort of 100 five year olds was to be tested four times over the span of the longitudinal study, it would be necessary to obtain four more groups of 100 five year olds, or, more practically, to subdivide the original 100 into five groups of 20 subjects each. Group I is tested at each testing session as in the usual longitudinal design, Group II is tested at time two and then discarded, Group III is also tested only once (time three), and Group IV is not tested until the fourth and final testing session. This design, and a possible statistical analysis, are given in Tables 3 and 4.

[Insert Tables 3 and 4 about here]

The ANOVA table for this design is admittedly rather complex. If the design is considered as a 2 x 4 x 4 factorial experiment with repeated measures on the least factor, then the ANOVA table becomes more obvious. (The three factors are: Practice - No Practice with 2 levels, 4 cohort groups within each level of P, and 4 age levels.) The problem is that there are repeated measures under P_1 but not under P_2, thus the difference among cohorts within levels of P are kept separate and different error terms are necessary to test these effects. Neither Baltes (1968) nor Schaie (1965) provide adequate descriptions of suitable statistical analyses for their designs. Baltes discusses it in a general way, and Schaie presents an ANOVA table for a complete factorial experiment with a randomized groups design. Failure to account for the repeated measures aspect of this design seems to be a serious flaw in Schaie's analysis.

It is interesting to note that the well known Solomon Four-Group design (Solomon, 1949; Solomon and Lessac, 1968) is very similar to these cross-sequential designs which control for the testing effect. The primary difference is that Solomon's designs are pre-post only, rather than longitudinal.

The Attrition Problem in Longitudinal Studies

A serious problem confronting all researchers involved in longitudinal studies is subject attrition, whether it is movers, resistors, or deceased subjects. The two main concerns of the investigator are; how can missing subjects be retrieved? and what statistical procedures are appropriate for repeated measures designs with incomplete data?

Retrieval procedures. McAllister, Butler, and Goe (1973) provide detailed procedures for relocating subjects in longitudinal studies. Their accompanying flow chart is a virtual recipe of step-by-step procedures. Their strategy was utilized in 1972 in an attempt to locate a random sample of 600 subjects from a sample of 2661 original participants in a 1963 survey. The 1963 sample consisted of 9 to 14 year olds, thus the 1972 sample ranged in age from 18 to 24 years - a very mobile group. Despite this, and the nine year time span, McAllister and his coworkers were able to trace over 90% of the 600 subjects. County marriage records, Postal Service back files, telephone directories, criss-cross directories, County Voter Registration files, school transfer records, Public Utilities Credit offices (which are considerably cheaper than the often recommended Retail Credit Unions), and State Departments of Motor Vehichles all proved to be useful information sources.

Statistical analysis. Attrition is not a serious problem in those designs which employ concomitant control groups. However, the majority of longitudinal studies presently underway probably are of the simple basic design, that is, a single group of individuals has been tested at time zero and then observed and tested at regular intervals for a number of years following. By the end of year five it is quite possible only 75% of the original sample remains, and, to further complicate the analyses, replacement subjects have been added in an attempt to retain a relatively stable sample size. Assuming that the investigation involves more than one dependent variable, and that the researcher

wishes to make statistical statements regarding the probability of significant changes while maintaining a relatively low experiment-wise error rate, then multivariate statistics are necessary, MANOVA being the most appropriate technique in most cases. Under these conditions there is only one option - delete from the statistical analysis all subjects for which there is not complete data. It does not matter if there are unequal numbers in the different groups (cohorts, or an a priori classification variable), but each subject must have a complete set of scores (i.e., each variable at each measurement period). It is as straightforward and unequivocal as that - delete all subjects with incomplete data. This applies only to the MANOVA analysis. There are a number of ways by which missing data can be replaced with estimators (i.e., Frane, 1976), but the basic assumption underlying all such methods is that the data are missing at random. As this is not the case in most longtitudinal studies, such procedures are invalid.

Additional valuable information can be gained by comparing the variable means at time zero for the partial-data subjects with the complete-data subjects. This, of course, tells nothing about development, but it does provide an indication of the extent to which the MANOVA results can be generalized to the initial population. The adding of subjects to longitudinal studies after the initial measures have been taken is certainly not recommended. As well as the problem of incomplete data, there are also problems related to differential testing effects, and selective sampling.

THE USE OF DIFFERENCE SCORES AS A MEASURE OF CHANGE

In a typical pretest-posttest repeated measures design, the resultant difference score, or gain score, is usually of primary interest to the researcher - despite its well known and frequently documented associated statistical problems. Objections to the use of difference scores have been made by methodologists for many years, were clearly defined by Bereiter (1963) approximately 15 years ago, and yet are still being made and debated today (Levin and Marascuilo, 1977). The following section examines different methods of computing criterion difference scores, and some possible adjustment procedures, and outlines the basic problems associated with the use of such scores.

Selection of a Criterion Score (unadjusted)

If the research methodology utilized yields a single score on the first administration of a test (X_1) and another single score on a repetition of that test at some subsequent point in time (X_2), then there is little choice in the criterion score to use if the researcher wishes to use a single, unadjusted, dependent variable. It has to be this difference ($D = X_2 - X_1$) - which has many inherent deficiencies and numerous possible transformations to reduce these deficiencies (none of which are very satisfactory). These are discussed later. A more likely situation, however, is when there are a number of observations available for each S (e.g., heart rate at each minute of a 15-minute exercise bout, 30 learning trials), but the investigator wishes to reduce this data to a single change score or learning score. The problems then confronting him are: (1) how many trials should he use to estimate both the initial and final states of the Ss?, and (2) should he use the best, or the average, of each of these sets of trials? Before commenting on some possible solutions to these two problems, it should be noted that neither of these problems should ever arise when dealing with the analysis of change. Discarding or reducing data, when suitable statistical methods are available for analyzing all available data, seams like very inefficient research. If the goal is the be able to understand motor behavior, for purposes of explanation and prediction, then one must look at all the data, and analyze it by a repeated measures ANOVA, time series, or some other equally suitable tool. However, many investigators insist on obtaining a single change score, thus some discussion on these points seems necessary.

The problem of choosing between the best and the average score has only one acceptable solution - use the average. There is sufficient support for use of the average rather than the best in the general case (Baumgartner, 1974; Henry, 1965; Kroll, 1967) and in the specific case of difference scores it is even more necessary. The reliability of a difference score is so dependent upon the reliability of the two scores which produce this difference, that it is imperative that these two scores possess maximum reliability themselves - thus averages are necessary.

The solution to the question of the optimal number of trials to use in computing these pre and post-score averages is not quite so unambiguous. The problem facing an investigator who uses a learning task is how can he choose a score which maximizes both reliability and discriminability at the same time? In a task which has, say, 20 trials, the difference between trial one and trial 20 will probably show the greatest discriminability as far as learning is concerned; however, it may not be very reliable. If one uses the average of the first ten trials as an indication of initial score, and the average of the last ten as the performance score, then the difference between these two may show high reliability, but it probably will not show much learning. Carron and Marteniuk (1970) pointed out the necessity for comparing the differences between both the reliabilities and discriminability obtained by grouping trials in different ways. Others (Baumgartner and Jackson, 1970; McCraw and McClenney, 1965) have attempted to give definitive rules for determining the number of trials and the measurement schedules one should employ. Because of the great variability in type of task, characteristics of Ss, etc., it does not seem possible to choose a specific rule for determining the "best" criterion measure for all situations - even for all situations involving a specific task or set of measures. If one decides that it is necessary to reduce the data to a single dependent variable (which, to this writer, does not seem to be a valid procedures), then utilizing procedures as suggested by Carron and Marteniuk (1970), and following the basic principles of reliability and validity of dependent variable scores which have been frequently and explicitly laid out for us (e.g.; Alexander, 1947; Burt, 1955; Feldt and McKee, 1957; Krause, 1969; Lomnicki, 1973; Schutz and Roy, 1973) one should be able to arrive at a procedure for selecting the most suitable criterion score in each specific situation.

Selection of a Criterion Score (adjusted)

In situations where there are only two opportunities for observation and measurement (pre and post), or where the investigator insists on reducing repeated measures to a pre-post case, then it is probably necessary to apply some type of statistical adjustment or correction factor to either the difference score or to the final score. The following section gives possible solutions for each of a number of common problems associated with using difference scores.

These problems have been well defined by many investigators (Bereiter, 1963; Cronbach and Furby, 1970; Lord, 1956, 1963; McNemar, 1958).

(i) Problem 1. Regression Effect: In general, on the second administration of a test, and in the absence of any true change or treatment effect, the observed scores for those who scored high on test #1 tend to decline and the observed scores of those who scored lowest on test #1 tend to increase on test #2.

Solutions. The most valid, and least complicated, solution, is to use a homogeneous group so all Ss have essentially the same initial score. If the experiment involves comparisons between groups, then equate the group means initially, either by randomization with large sample sizes, blocking, matching, or statistically through analysis of covariance.

Another possible solution, the one to which psychometricians have directed their attention, is to adjust the final score on the basis of the pre-post linear regression effect. This can be done by fitting a regression line to the pre-post scores (X_1, X_2) under the conditions of the null hypothesis; i.e., no treatment effect, and then use deviation from the regression line as the dependent variable indicating true change (Lord, 1963). This requires either a separate control group or a (X_1, X_2) measure for each subject under a treatment condition and a control condition - a procedure which is not always possible. The most reasonable solution seems to be to use analysis of covariance (ANCOVA) as it is essentially an analysis of the X_2 scores, adjusted on the basis of the regression line between X_2 and X_1.

(ii) Problem 2. Measurement Errors or the Unreliability-Invalidity Dilemma: The degree to which measurement errors exist in the initial and/or final measures, along with the degree to which the X_1, X_2 correlation exceeds zero, is reflected by a reduction in the reliability of the X_1-X_2 difference score.

Solutions. There exists a wealth of information on possible solutions to this problem (e.g.; Lord, 1956, 1963; McNemar, 1958; Ng, 1974; Tucker, 1966; Wiley and Wiley, 1974).

The basic thesis of all these articles is that it is possible to compute a reliability coefficient 'corrected for attenuation', that is, the reliability of a difference between 'true scores' (errorless measures yielding reliabilities of 1.00 in both X_1 and in X_2). Once having obtained a reliable estimate of true difference it is then possible to use this attenuated reliability coefficient and multiply it by the observed X_2-X_1 difference (but scaled as deviations from the means), thus obtaining a hypothetical true difference score or "regressed score" (McNemar, 1958). Although this is the basis of the solutions advocated by many psychometricians it has its deficiencies, the primary one being that the number of alternate ways to compute this true gain score seems to be exceeded only by the number of papers written on the topic. The non-specialist is left with a morass of equations and confusion. Another deficiency with the use of estimated true difference scores is that the regression coefficient used in the predictor equation is based on a number of assumptions, some of which may not always hold true. A recent report by Wiley and Wiley (1974) indicates that the assumption of independence of errors of measurement between tests is frequently violated, thus giving overestimates of the attenuated reliability coefficient. This in turn would result in overestimates of the true gain score.

(iii) Problem 3. Equality of Scale Along the Range of Scores (the Physicalism-Subjectivism Dilemma): An observed score at the low range of the continuum may be measuring an attribute of behavior quite different from that which is reflected by the same test at the high end of the range of scores.

Solutions: There seem to be no adequate solutions per se for this problem. One could use P-technique methodology (a sort of factor analysis appropriate for change data) to test the assumption that the two measures are in fact measuring the same thing (Bereiter, 1963; Cattell, 1963). However, this is not a solution, but rather a technique to reveal the existence or non-existence of a problem. The answer seems to be in finding ways to avoid the problem rather than solve it - and this can be accomplished to a limited degree. If all groups are equated initially with respect to their scores on the dependent variable, then any differences between groups in the amount of change within groups can be logically interpreted (Schmidt, 1972).

This restriction allows for the conclusion that one group changed more, or less, with regards to the particular dependent variable being used. If one group showed very large changes, and the other group very small ones, then it may be difficult to interpret the meaning of the relative magnitudes of change scores, but it is still possible to state that one group should significantly greater change than the other group on that particular trait.

A General Solution to the Problems Associated with Difference Scores

At this point the reader must be wondering, "Is there no adequate solution to the problem of measuring change?" My answer is "Yes" there are adequate methods, but not through the use of difference scores. If one must use a change score, then perhaps the "best" estimator of a true difference score is Cronbach and Furby's "complete estimator" (1970):

$$\hat{D}_{\infty} = \hat{X}_{1_{\infty}} - \hat{X}_{2_{\infty}}$$

where $\hat{D}_{\infty}$ is the "true difference score", and $\hat{X}_{1_{\infty}}$ is the true score at time 1, taking into account numerous other categories of variables, W, which may be multivariate in nature and relate to the pre or post scores in some manner. The true score for X_1 is estimated as:

$$\hat{X}_{\infty} = pxx'X_1 + \frac{\sigma X1_{\infty}(X_2 \cdot X_1)}{\sigma^2(X_2 \cdot X_1)} (X_2 \cdot X_1) + \frac{\sigma X1_{\infty}(W \cdot X_1, X_2)}{\sigma^2(W \cdot X_1, X_2)} (W \cdot X_1, X_2) + \text{constant}$$

where $(X_2 \cdot X_1)$ and $(W \cdot X_1, Y_1)$ are partial variates. The purpose of presenting this equation is not to provide the reader with a useful statistical tool, but rather to point out the extreme degree to which the raw data can be transformed if one wishes some sort of pure measure. The difficulty in interpreting this transformed score is obvious - at least in terms of predictable observed behavior.

Two quotes provide a suitable summary of this investigator's position on the use of difference scores:

> "Both the history of the problem and the logic of investigation indicate that the last thing one wants to do is think in terms of or compute such change scores unless the problem makes it absolutely necessary." (Nunnally, 1973, p. 87)

> "Gain scores are rarely useful, no matter how they may be adjusted or refined." (Cronbach and Furby, 1970, p. 68)

The Statistical Analysis of Difference Scores

Given a single group, pretest-posttest design, there are two equivalent ways to test the null hypothesis of a zero mean difference, namely a t test for correlated means or a one-way repeated measures ANOVA (the F ratio of the ANOVA will be identical to t^2). Of concern here are the consequences of the unreliability of the difference scores.

The measurement specialist is primarily concerned about reliability as a phenomenon in itself, placing high value on reliabilities near 1.0 and showing abhorrence at values of less than .50. Assuming that the reliabilities of the pretest and posttest are the same (r) and given the correlation between pretest and posttest as r_{12}, then the reliability (r_d) of the difference score is:

$$r_d = \frac{r - r_{12}}{1 - r_{12}}$$

Thus as r_{12} approaches r, the reliability of the difference score approaches zero. In order to attain a high r_d, the magnitude of r_{12} must be small related to r; i.e., if $r_{12} = .25$, $r = .75$, then $r_d = .67$. However, this does little to appease the measurement specialist as an r_{12} of .25 suggests that the test is not measuring the same attribute at each point in time. Consequently, the researcher either avoids difference scores or attempts to "correct" them as discussed earlier in this paper.

The statistician on the other hand views low reliability in difference scores with fewer misgivings, because as this reliability decreases, the power of the statistical test increases. As is shown above, it is the value of r_{12} which is of importance (for a fixed value of r). This can be demonstrated in both the ANOVA and t tests. In the latter, the denominator $S_{\bar{D}}$ approaches zero as r_{12} approaches 1.0, thus minimizing the denominator and maximizing the calculated t. For a common variance ($S_1^2 = S_2^2 = S^2$) and $r_{12} = 1.0$;

$$S^2_{\bar{D}} = \frac{S_1^2 + S_2^2 - 2r_{12} S_1 S_2}{n} = \frac{2S^2 - 2S^2}{n} = 0.0$$

Similarly for the F ratio in ANOVA. As r_{12} approaches one, the Subjects by Trials interaction approaches zero, thus maximizing the F ratio for the Trials effect.

Thus, although the reliability of the tests themselves should be important to researchers, the reliability of the difference scores may not be that crucial.

UNIVARIATE ANOVA AND MANOVA

The analysis of all of the available data should provide an investigator with more information than does the limited, and suspect, information provided in a difference score. These repeated measures analyses may be performed by either univariate or multivariate analysis of variance (ANOVA, MANOVA) on the raw scores or on scores adjusted for initial differences between groups. The more information available on the nature of change in behavior over time, the greater should be the degree of understanding of the nature and causes of that change. Consequently, in an experiment involving any length of time between the initiation of the treatment and the final observation, it is desirable to take numerous measures per S. Although in some cases it is not possible to do this, either due to the contamination effect of the measurement tool or to the nature of the treatment procedures, in most motor behavior studies such repeated measures are quite feasible.

Repeated Measures ANOVA

The common method for analyzing change for a repeated measures design is through a repeated measures or Ss x Treatments ANOVA. Given a typical experiment involving two treatment groups (or a treatment and control) with 20 Ss nested within each group and repeated across say 10 trials (Fig. 1), one appropriate method for analyzing change could be to break down the total variability as given in Table 5.

[Insert Fig. 1 and Table 5 about here]

The effects of most interest here, with respect to the analysis of change, are the Groups x Trials and its trend analysis components, Groups x Trials (Linear) and Groups x Trials (Quadratic). The Groups x Trials interaction indicates the degree to which the change over trials is the same for each group - which is probably the research question of most interest; i.e., is there a significant change in behavior over the time span of the experiment, and, if so, does this change show the same, or different, characteristics between the two experimental groups?

The Groups x Trials (Linear) asks essentially the same question but with the constraint that the change over time is linear. In this case a linear function is forced on the data and the test of significance tests for equality of slopes between the two groups, which in behavioral terms amounts to a comparison of the rates of learning, rates of recovery, etc. Similarly the Groups x Trials (Quadratic) compares the two treatment groups on the basis of the degree of curvature or time of plateauing of the scores over time.

This analysis then provides one possible solution for the analysis of change suitable for many experimental conditions. By using a number of measures instead of just two, the problems of regression effect and measurement errors are greatly reduced. The unreliability of the data is reflected by the magnitude of the S x Trials interaction (or in this case the S(G) x T) and is thus a sort of built in protection against making erroneous research conclusions based on unreliable data. The less reliable the data is, the larger the S x Trials error term, the more difficult it is to attain statistical significance and the less likely it is to make a Type I error.

The repeated measures ANOVA is not the ideal solution to the problems of analyzing change, however, for a number of reasons. Firstly, the tests of significance give limited information regarding the nature or form of the change over time, as the trend analyses fit only polynomials to the data, data which is frequently better fitted by a logarithmic or exponential function. Secondly, it deals with mean values only and does not reveal reliable differences between subjects (within the same group) with respect in intra-individual behavioral changes over time (a stochastic model would detect this). Finally, and perhaps most importantly, the nature of the data common to most studies in motor behavior is such that it violates the assumptions on which the repeated measures ANOVA is founded. These assumptions are that the measures (i) are normally distributed, (ii) exhibit equal variances under all treatment conditions, and (iii) have equal covariances between all treatment pairs (the precise mathematical assumption is that all covariances equal zero but the F ratio is virtually unaffected by violation of this assumption, providing all covariances are equal). While the first two of these assumptions are usually met with motor performance data, the third one rarely is.

This assumption can be casually tested by examining the correlation matrix of the repeated measures - the degree to which all correlations are not equal indicates the degree to which this assumption is violated.[1] It is frequently the case in our field of study to obtain data in which adjacent trial correlations are very high, but diminish as a function the number of intervening observations between any two measures. The resultant of this situation is an inflated F value and a substantial increase in the probability of committing a Type I error (as high as $p = .15$ when assuming $\alpha\ p = .05$).

The analysis of variance for repeated measures, which was first presented here as a possible solution to some of the problems inherent in the analysis of change, has now become a problem itself. There are two possible ways by which ANOVA may be validly used on repeated measures data which exhibits unequal between trial correlations:

(1) Inflate the magnitude of the F needed for significance by reducing the associated degrees of freedom (d.f.). Box (1954) has suggested that the d.f. for both the numerator and denominator be multiplied by a factor ε, which is a function of the degree of heterogeneity of both the variances and the covariances. The greater the heterogeneity the smaller the calculated ε and the larger the F value must be in order to reject the null hypotheses.

(2) Greenhouse and Geisser (1959) questioned the validity of the estimator ε and its effect on the approximate F distribution. They suggested the use of the minimum possible value of ε, namely $1/(k-1)$ where k is the number of levels of the repeated factor, as the factor which should be applied to the d.f. in all situations. Although this is a statistically valid technique it is very conservative, thus resulting in a rather large probability of committing a Type II error.

There are a number of excellent articles available which provide a lucid explanation of both the problem and the merits of these solutions (e.g., Davidson, 1972; Gaito, 1973; Gaito and Wiley, 1963; McCall and Appelbaum, 1973; Mendoza, Toothaker and Nicewander, 1974).

---

[1]Procedures for statistical tests of this assumption are available in Winer (1971, p. 594).

Repeated Measures MANOVA

The other solution to the problem of non-homogeneity of covariances is to use a technique which does not require this assumption - namely the multivariate analysis of variance. MANOVA requires no assumptions regarding the homogeneity of covariances and allows for an exact statistical test based on a known significance level. Although this technique has been available for many years, it has not been adopted by practicing researchers due to its extreme computational complexity. However, the present accessibility of suitable computerized multivariate statistical packages at most universities has eliminated such an excuse for ignoring this very useful test and it should now be a standard statistical tool for all researchers. Very briefly, what MANOVA does is to transform the k repeated measures for each subject into a set of (k-1) scores through the application of independent contrasts (these are usually orthogonal polynomials, but they need not be as the resulting significance test is independent of the choice of contrasts). An analysis of variance type procedure is then carried out on the vector of means of these derived scores with the mean square error being a variance-covariance matrix of within cell variabilities rather than a unitary scalar value as in the univariate procedure. The tests of significance provide an F ratio for the overall multivariate hypothesis that the trial means are equal, and for a two group experiment, that the change in performance across repeated measures is the same for each group. An overall significant F on these multivariate hypotheses allows the investigator to use appropriate follow-up tests while maintaining an overall pre-determined level of significance. These follow-up procedures can take the form of simultaneous confidence intervals, step-down F ratios, or even the usual univariate F tests on each dependent variable separately or on the single d.f. contrasts associated with trend analysis (see Spector, 1977, for a good review of procedures).

Another frequently used procedure associated with MANOVA is discriminant analysis which tests whether two or more groups can be significantly separated on the bases of their profiles (or, in the RM design, their pattern of change over time). It has been shown, however, that a Groups x Trials ANOVA is more versatile in detecting the nature of the differences between group profiles than is discriminant analysis (Thomas and Chissom, 1973). Although Thomas and Chissom failed to consider the restrictive assumption inherent in the univariate G x T ANOVA, this is not a factor if the Trials effect is broken down into polynomial coefficients (linear, quadratic, etc.).

This essentially converts the univariate procedure to a multivariate technique and thus no longer requires the assumption of equal covariances. Bock (1963), Cole and Grizzle (1966), and Finn (1969) have provided comprehensive discussions on the application of MANOVA to repeated measures data, and comparisons of the applications and outcomes of ANOVA versus MANOVA are well given by Davidson (1972), Hummel and Sligo (1971), McCall and Appelbaum (1973), and Poor (1973).

It must be pointed out that not all statisticians nor psychometricians favor multivariate methods. Kempthorne (1966, p. 521) has stated that, "I have yet to see any convincing examples of experimental data in which the standard techniques of multivariate analysis have led to scientific insight." Perhaps the choice between these two types of analyses can be based on whether the experimental study is primarily concerned with "information finding" or with "decision making". Univariate procedures may allow for greater (or easier) interpretation of the data, and thus support the information finding approach, whereas multivariate techniques (MANOVA in particular), by providing an exact probability statement, are most suitable for decision making.

Table 6 provides for a comparison of the relative powers of multivariate (MV) and univariate (UV) F tests. The symbol MUV refers to a repeated measures univariate ANOVA which has been modified by the Greenhouse and Geisser technique. Notice that for small N the MV procedure lacks power in all cases. For large N (20 more than the number of dependent measures) and a relatively large non-centrality parameter (δ), the slightly greater power of the UV over the MV procedure is more than compensated for by the lower experimentwise error rates in the MV methods. In these situations a MANOVA would seem preferable to an ANOVA.

CONCLUSION

There are obviously a considerable number of problems inherent in the measurement and analyses of change, especially in research designs of a longitudinal nature. However, most of these problems can be avoided provided sufficient care and planning are taken prior to initiating the research project. The cross-sectional sequential type designs which are required for valid measures of developmental change are very costly - but necessary if the research is to have any scientific value. Multivariate statistical procedures utilizing on complete data sets will provide for valid and relatively powerful tests of hypotheses.

		Trials			
		T_1	T_2	. . .	T_{10}
Group 1	S_1	X_{111}	X_{112}	. . .	X_{1110}
	S_2	X_{121}	X_{122}	. . .	X_{1210}
	.	.	.		.
	.	.	.		.
	.	.	.		.
	S_{20}	X_{1201}	X_{1202}	. . .	X_{12010}
Group 2	S_1	X_{211}	X_{212}	. . .	X_{2110}
	S_2	.	.		.
	.	.	.		.
	.	.	.		.
	.	.	.		.
	S_{20}	.	.		X_{22010}

Fig. 1. Schemata of 2 x 10 Factorial Experiment with Repeated Measures on the Second Factor.

TABLE 1

SEQUENTIAL RESEARCH DESIGN GIVING AGES OF COHORT GROUPS AT EACH TESTING TIME

Cohort	Time of Measurement						
	1930	1940	1950	1960	1970	1980	1990
1930	5	15	25	35			
1940		5	15	25	35		
1950			5	15	25	35	
1960				5	15	25	35

TABLE 2

ANALYSIS OF VARIANCE FOR A p x q BIFACTOR DEVELOPMENTAL DESIGN

Source of Variation	df	Mean Square	F
Among Cohorts (C)	p-1	MS_C	MS_C/MS_{SwC}
Ss within Cohorts (SwC)	p(n-1)	MS_{SwC}	
Age (A)	q-1	MS_A	MS_A/MS_{SwCA}
Cohort x Age (CA)	(p-1)(q-1)	MS_{CA}	MS_{CA}/MS_{SwCA}
SwCohort x Age (SwCA)	p(n-1)(q-1)	MS_{SwCA}	

TABLE 3

EXPERIMENTAL LAYOUT FOR A BIFACTOR DEVELOPMENTAL DESIGN WITH CONTROL GROUPS FOR TESTING EFFECTS

Cohort	Age at Time of Testing			
	5	15	25	35
1930 (S_{1-20})	✓	✓	✓	✓
(S_{21-40})	✓	X	X	X
(S_{41-60})	X	✓	X	X
(S_{61-80})	X	X	✓	X
(S_{81-100})	X	X	X	✓
1940 (S_{1-20})	✓	✓	✓	✓
.				
.				
.				
(S_{81-100})	X	X	X	✓
.				
.				
.				
1960 (S_{1-20})	✓	✓	✓	✓
.				
.				
.				
(S_{81-100})	X	X	X	✓

X - denotes no testing at this time.
✓ - denotes testing done at this time.

TABLE 4

ANALYSIS OF VARIANCE FOR A BIFACTOR DEVELOPMENTAL DESIGN WITH CONTROL GROUPS FOR TESTING EFFECTS

Source of Variation	df[1]		
Practice Effects (P)	1		
Cohorts with P_1 (CwP_1)	3		
Subjects with CwP_1 ($SwCwP_1$)		76	- error term for CwP_1
Cohorts within P_2 (CwP_2)	3		
Within Cell in P_2 (Error P_2)		304	- error term for CwP_2
Error for P (Error P)		380	- pooled Error P_2 and $SwCwP_1$
Age (A)	3		
A x P	3		
A x CwP_1	9		
A x CwP_2	9		
$SwCwP_1$ x A		228	- error for A x CwP_1
Pooled Error	608		- error P + $SwCwP_1$ x A - error term for A, A x P, A x CwP_2
Total	639		

[1] The degrees of freedom are based on the design given in Table 3.

TABLE 5

Analysis of Variance, with Trend, for a 2 x 10 Factorial Experiment with Repeated Measures on the Second Factor

Source	df	Mean Square	F Ratio
Groups	1	MS_G	$MS_G/MS_{S(G)}$
Ss within Groups	38	$MS_{S(G)}$	
Trials	9	MS_T	$MS_T/MS_{S(G)T}$
Linear	1	MS_{T_L}	$MS_{T_L}/MS_{S(G)T}$
Quadratic	1	MS_{T_Q}	$MS_{T_Q}/MS_{S(G)T}$
Residual	7	MS_{T_R}	$MS_{T_R}/MS_{S(G)T}$
Groups x Trials	9	MS_{GT}	$MS_{GT}/MS_{S(G)T}$
G x $T_{Lin.}$	1	MS_{GT_L}	$MS_{GT_L}/MS_{S(G)T}$
G x $T_{Quad.}$	1	MS_{GT_Q}	$MS_{GT_Q}/MS_{S(G)T}$
G x $T_{Resid.}$	7	MS_{GT_R}	$MS_{GT_R}/MS_{S(G)T}$
SwG x Trials	342	$MS_{S(G)T}$	
Total	399		

Table 6.

Relative Power of Multivariate and Univariate F Tests

No. of Trials	$\delta/\sqrt{k}$	F Test	Equal var.-cov.?	Power	
				n=k+1	n=k+20
3	1.0	UV	Yes	.21	.30
		MUV	Yes	.07	.18
		MUV	No *	.23	.38
		MV	-	.12	.28
3	2.0	UV	Yes	.66	.86
		MUV	Yes	.34	.75
		MUV	No	.64	.91
		MV	-	.30	.83
6	1.0	UV	Yes	.39	.45
		MUV	Yes	.03	.07
		MUV	No	.54	.65
		MV	-	.11	.34
6	2.0	UV	Yes	.97	.99
		MUV	Yes	.46	.76
		MUV	No	.98	.99
		MV	-	.26	.93

* For a specific case of non-uniform variance-covariance matrix

References

Alexander, H. W. The estimation of reliability when several trials are available. Psychometrika, 1974, 12, 79-99.

Baltes, P. B. Longitudinal and cross-sectional sequences in the study of age and generation effects. Human Development, 1968, 11, 145-171.

Baumgartner, T. A. Criterion score for multiple trial measures. Research Quarterly, 1974, 45, 193-198.

Baumgartner, T. A., & Jackson, A. S. Measurement schedules for tests of motor performance. Research Quarterly, 1970, 41, 10-14.

Bentler, P. M. Assessment of developmental factor change at the individual and group level. In J. R. Nesselroade, & H. W. Reese (Eds.), Life-span developmental psychology. New York: Academic Press, 1973.

Bereiter, C. Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), Problems in measuring change. Madison: University of Wisconsin Press, 1963.

Bock, D. Multivariate analysis of variance of repeated measurements. In C. W. Harris (Ed.), Problems in measuring change. Madison: University of Wisconsin Press, 1963.

Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems: II. Effects of inequality of variance and correlation between errors in the two way classification. Annals of Mathematical Statistics, 1954, 25, 484-498.

Burt, C. Test reliability estimated by analysis of variance. British Journal of Statistical Psychology, 1955, 8, 103-118.

Carron, A. V., & Marteniuk R. G. An examination of the selection of criterion scores for the study of motor learning and retention. Journal of Motor Behavior, 1970, 2, 239-244.

Cattell, R. B. The structuring of change by P-technique and incremental R-technique. In C. W. Harris (Ed.), Problems in measuring change. Madison: University of Wisconsin Press, 1963.

Cole, J. W. L., & Grizzle, J. E. Applications of multivariate analysis of variance to repeated measurements experiments. Biometrics, 1966, 22, 810-828.

Cronbach, L. J., & Furby, L. How we should measure "change" or should we? Psychological Bulletin, 1970, 74, 68-80.

Davidson, M. L. Univariate versus multivariate tests in repeated measures experiments. Psychological Bulletin, 1972, 77, 446-452.

Feldt, L. S., & McKee, M. E. Estimation of the reliability of skill tests. Research Quarterly, 1957, 29, 279-293.

Finn, J. D. Multivariate analysis of repeated measures data. Multivariate Behavioral Research, 1969, 4, 391-413.

Frane, J. W. Some simple procedures for handling missing data in multivariate analysis. Psychometrika, 1976, 41, 409-415.

Gaito, J. Repeated measurements designs and tests of null hypotheses. Educational and Psychological Measurement, 1973, 33, 69-75.

Gaito, J., & Wiley, D. E. Univariate analysis of variance procedures in the measurement of change. In C. W. Harris (Ed.), Problems in measuring change. Madison: University of Wisconsin Press, 1963.

Gottman, J. M., McFall, R. M., & Barnett, J. T. Design and analysis of research using time series. Psychological Bulletin, 1969, 72, 299-306.

Greenhouse, S. W., & Geisser, S. On methods in the analysis of profile data. Psychometrika, 1959, 24, 95-111.

Henry, F. M. "Best" versus "Average" individual scores. Research Quarterly, 1965, 38, 317-320.

Huck, S. W. & McLean, R. A. Using a repeated measures ANOVA to analyze the data from a pretest-posttest design. A potentially confusing task Psychological Bulletin, 1975, 82, 511-518.

Hummel, T. J., & Sligo, J. R. Empirical comparison of univariate and multivariate analysis of variance procedures. Psychological Bulletin, 1971, 76, 49-57.

Kempthorne, O. Multivariate responses in comparative experiments. In P. R. Krishnaiah (Ed.), Multivariate analysis. New York: Academic Press, 1966.

Krause, M. S. The theory of measurement reliability. Journal of General Psychology, 1969, 80, 267-278.

Kroll, W. Reliability theory and research decision in selection of a criterion score. Research Quarterly, 1967, 38, 412-419.

Labouvie, E. W., Bartsch, T. W., Nesselroade, J. R., & Baltes, P. B. On the internal and external validity of simple longitudinal designs. Child Development, 1974, 45, 282-290.

Levin, J. R., & Marascuilo, L. A. Post hoc analysis of repeated measures interactions and gain scores: Whither the inconsistency? Psychological Bulletin, 1977, 84, 247-248.

Lomnicki, Z. A. Some aspects of the statistical approach to reliability. Journal of Royal Statist. Soc. A., 1973, 136, 395-419.

Lord, F. M. The measurement of growth. Educational and Psychological Measurement, 1956, 16, 421-437.

Lord, F. M. Elementary models for measuring change. In C. W. Harris (Ed.), Problems in measuring change. Madison: University of Wisconsin Press, 1963.

McAllister, R. J., Butler, E. W., & Goe, S. J. Evolution of a strategy for for the retrieval of cases in longitudinal survey research. Sociology and Social Research, 1973, 58, 37-47.

McCall, R.B., & Appelbaum, M. I. Bias in the analysis of repeated measures designs: Some alternative approaches. Child Development, 1973, 44, 401-415.

McGraw, L. W. & McClenney, B. N. Reliability of fitness strength tests. Research Quarterly, 1965, 36, 289-295.

McNemar, Q. On growth measurement. Educational and Psychological Measurement, 1958, 18, 47-55.

Mendoza, J. L., Toothaker, L. E., & Nicewander, W. A. A Monte Carlo comparison of the univariate and multivariate methods for the groups by trials repeated measures design. Multivariate Behavioral Research, 1974, 9, 165-177.

Ng, K. T. Applicability of classical test score models to repeated performances on the same test. Australian Journal of Psychology, 1974, 26, 1-8.

Nunnally, J. C. Psychometric Theory. New York: McGraw Hill, 1967.

Nunnally, J. C. Research strategies and measurement methods for investigating human development. In J. R. Nesselroade, & H. W. Reese (Eds.), Life-span developmental psychology. New York: Academic Press, 1973.

Overall, J. E., & Woodward, J. A. The unreliability of difference scores: A paradox for measurement of change. Psychological Bulletin, 1975, 82, 85-86.

Poor, D. D. S. Analysis of variance for repeated measures designs: Two approaches. Psychological Bulletin, 1973, 30, 204-209 .

Rose, C. L. Representativeness of volunteer subjects in a longitudinal aging study. Human Development, 1965, 8, 152-156.

Schaie, K. W. A general model for the study of developmental problems. Psychological Bulletin, 1965, 64, 92-107.

Schaie, K. W. Methodological problems in descriptive developmental research on adulthood and aging. In J. R. Nesselroade, & H. W. Reese (Eds.), Life-span developmental psychology: Methodological issues. New York: Academic Press, 1973.

Schaie, K. W., LaBouvie, G. V., & Barrett, T. J. Selective attrition effects in a fourteen-year study of adult intelligence. Journal of Gerentology, 1973, 28, 328-334.

Schaie, K. W., & Strother, C. R. A cross-sectional study of age changes in cognitive behavior. Psychological Bulletin, 1968, 70, 671-680.

Schmidt, R. A. The case against learning and forgetting scores. Journal of Motor Behavior, 1972, 4, 79-88.

Schutz, R. W. Stochastic processes: Their nature and use in the study of sport and physical activity. Research Quarterly, 1970, 41, 205-212.

Schutz, R. W., & Roy, E. A. Absolute error: The devil in disguise. Journal of Motor Behavior, 1973, 5, 141-153.

Solomon, R. L. An extension of control group design. Psychological Bulletin, 1949, 46, 137-150.

Solomon, R. L., & Lessac, M. S. A control group design for experimental studies of developmental processes. Psychological Bulletin, 1968, 70, 145-150.

Spector, P. E. What to do with significant multivariate effects in multivariate analyses of variance. Journal of Applied Psychology, 1977, 62, 158-163.

Terman, L. M. Genetic studies of genius: Mental and physical traits of a thousand gifted children, Vol. I. Stanford: Stanford University Press, 1925.

Thomas, J. R., & Chissom, B. S. Comparison of groups x trials analysis of variance and discriminant analysis for use in group profile evaluations. Perceptual and Motor Skills, 1973, 37, 671-675.

Tucker, L. R., Damarin, F., & Messick, S. A base-free measure of change. Psychomtrika, 1966, 31, 457-473.

Wall, W. D., & Williams H. L. Longitudinal studies and the social sciences. London: Heinemann, 1970.

Wiley, J. A., & Wiley, M. G. A note or correlated errors in [illegible]ated measurements. Sociological Methods and Research, 1974, 3, 172-188.